An experimental study of methods for the analysis of variance components in the inference of laboratory information

Anders Kallner & Elvar Theodorsson

To cite this article: Anders Kallner & Elvar Theodorsson (2020) An experimental study of methods for the analysis of variance components in the inference of laboratory information, Scandinavian Journal of Clinical and Laboratory Investigation, 80:1, 73-80, DOI: 10.1080/00365513.2019.1700426

To link to this article: https://doi.org/10.1080/00365513.2019.1700426
An experimental study of methods for the analysis of variance components in the inference of laboratory information

Anders Kallner and Elvar Theodorsson

Department of Clinical Chemistry, Karolinska University Hospital, Stockholm, Sweden; Department of Clinical Chemistry and Department of Clinical and Experimental Medicine, Linköping University, Linköping, Sweden

ABSTRACT
Measurement uncertainty (MU) can be estimated and calculated by different procedures, representing different aspects and intended use. It is appropriate to distinguish between uncertainty determined under repeatability and reproducibility conditions, and to distinguish causes of variation using analysis of variance components. The intra-laboratory MU is frequently determined by repeated measurements of control material(s) of one or several concentrations during a prolonged period of time. We demonstrate, based on experimental results, how such results can be used to identify the repeatability, ‘pure’ reproducibility and intra-laboratory variance as the sum of the two. Native patient material was used to establish repeatability using the Dahlberg formula for random differences between measurements and an expanded Dahlberg formula if a non-random difference, e.g. bias, was expected. Repeatability and reproducibility have different clinical relevance in intensive care compared to monitoring treatment of chronic diseases, comparison with reference intervals or screening.

Introduction
Determination and monitoring of quality indicators are crucial and mandatory for measurements in laboratory medicine. The indicators which are primarily used are based on internal quality control (IQC) and external quality assessment/proficiency testing (EQA/PT), both relying on regular measurements of control materials. The frequency of control measurements is regulated by international or national standards and stated in the Standard Operating Procedures of accredited quality systems. Professional groups have discussed and outlined how quality goals [1,2] should be established in laboratory medicine to best serve clinical needs.

The typical IQC program consists of measuring control materials of known – or assigned – concentrations a certain times per day and assuming that the measurements of patient samples performed between control measurements are acceptable. Typically, control materials at two concentration levels are used. There is an abundance of commercial control materials available and critical properties of these are commutability with patient samples in the used measuring system, and stability over time. Alternatively, patient material can be used to investigate performance using duplicate measurements [3,4]. Using the latter approach, the entire measuring interval can be covered using native samples. A combination of different approaches and algorithms will provide comprehensive information about the performance of the method.

Quality indicators are designed to monitor the trueness and the precision of the measurement results and the quality systems designed to ensure that results are consistent, comparable over time and transferable between service providers. Laboratories need to have detailed information on trueness and precision whereas the clinician, assuming a sufficient analytical quality, appropriately will focus on the transferability and comparability of results to reliably identify differences between results and between patient values and reference values. In the present study, we compare various procedures using control materials and patient samples to estimate measurement uncertainty, compare the usefulness of the results and use simulations to estimate the number of necessary duplicate samples for repeatability estimates.

Methods and materials

Measuring system
All measurements were using Roche Diagnostics (Germany) Cobas c501 employing reagents from Roche ALTL (Alanine Aminotransferase PyP) Ref 04467388190, ALP2S (ALP IFCC Gen.2) Ref 03333752190 and CA2 (Calcium Gen. 2) Ref 05061482190.

Stabilized control materials
The low and high concentrations reference materials were Autonorm Human Liquid L-1 1606342 and Autonorm Human Liquid L-2 1608804 (Sero A/S Oslo, Norway), respectively.
These materials are based on human serum with no preservatives or stabilizers added, for maximum commutability.

**Patient materials**

De-identified patient samples from the laboratory routine production were used. The selected analytes were P-ALT (Plasma-Alanine transferase), P-ALP (P-Alkaline phosphatase) and P-Calcium. The Swedish law of biobanking (SFS 2002:297) allows the use of anonymized samples for quality development and management. The results of the present study can never be traced to any individual or group of individuals. Accordingly, the study did not require scrutiny by the Ethics committee.

**Experimental design**

All calculations and estimations were based on the following experimental designs (Figure 1).

1. Control material of two concentrations were measured five times daily for five days and the ‘analysis of variance components’ (ANOC) was carried out in a $5 \times 5$ experimental design according to Clinical and Laboratory Standards Institute (CLSI) EP15 [4,5] in a specially developed MS Excel program [6]. This allows identifying the repeatability (within series variation), the ‘pure’ between series variation and the total variation (Figure 2).

2. Control materials of two concentrations were measured repeatedly, up to three times per day, totally 83 measurements of each quantity. The average, standard deviation ($s$) and %CV were calculated using standard procedures (Table 2).

3. The standard deviation was calculated from duplicate measurements of 100 consecutive patient samples, using the Dahlberg formula. The results of the expanded Dahlberg formula (ExpD) [4] were also calculated (Table 3).

4. The standard deviation ($s_D$) and the confidence intervals were calculated using the available

![Figure 1](image1.png)

**Figure 1.** The components and design of the study.

![Figure 2](image2.png)

**Figure 2.** Within and between series variations based on series of five measurements on five occasions (days). One of the results (as shown in the figure) in the first series was regarded as an outlier and not used for further calculations.
‘100-series’ of P-ALT, P-ALP and P-Calcium (P-Ca) (Figure 3), right panel. In creating these graphs, differences outside the average difference ±3 standard deviations of the differences were disregarded as not representative of the sample population.

All calculations and data handling were performed using EXCEL 2016.

Calculations

The analysis of variance components (ANOC) requires an experimental design in which observational results can be identified as within-groups and between-groups. This is also the design of an one-way ANOVA which yields the sum of squares and mean squares (MS) of the observations within- and between groups [7,8]. We used a model with five measurements within the groups, i.e. obtained under repeatability conditions and carried out on five occasions, i.e. under reproducibility conditions. This is summarized as a 5 × 5 table. The within-group mean square (MSw) is equal to the variance of the measurements within a group. The between group mean square (MSb) is a combination of the ‘pure’ between group variance and the within group variance. The pure between group variance \( s_b^2 \) was obtained by

\[
 s_b^2 = \frac{MS_b - MS_w}{n_0}
\]  

(1)

where \( n_0 \) is the average number of observations the groups. This is considered applicable in normal laboratory work and a reasonably balanced study.

The total or intra-laboratory variance (\( s_{inw}^2 \)) is

\[
 s_{tot}^2 = s_b^2 + MS_w
\]  

(2)

If \( \frac{MS_b - MS_w}{n_0} < 0 \), i.e. \( MS_b < MS_w \) then \( s_{tot}^2 = MS_w \), i.e. the within series variance [4].

The ANOC was also applied to the IQC data, assuming repeatability within days and reproducibility between days.

Imprecision was calculated from repeated measurements of control materials by the general formula for standard deviation (Equation (3)); the Dahlberg formula [3,4] (Equation (4)) was used for calculation of the repeatability assuming only a random variation between the original and replicate measurement and an expanded Dahlberg (ExpD) formula (Equation (5)) [4] when a non-random variation between the results was assumed.

\[
s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n - 1}
\]  

(3)

\[
s^2 = \frac{\sum_{i=1}^{n}d_i^2}{2 \times N}
\]  

(4)

\[
s^2 = \frac{\sum_{i=1}^{N}(d_i - \bar{d})^2}{2 \times (N - 1)}
\]  

(5)

where \( s^2 \) is the variance, \( n \) is the number of observations, \( N \) is the number of pairs in a comparison of duplicates and \( d \) is the difference between duplicates.

Simulations

Five different situations were identified in which duplicate measurements may be used for estimating the repeatability of a measurement procedure. In all examples, the bias was assumed to be negligible and the expanded Dahlberg results therefore not calculated. These simulations were undertaken to determine the minimal number of observations to achieve a reliable estimate of the Dahlberg standard deviation.

1. The original and replicate measurements were simulated by adding a random difference to a simulated normally distributed series of results with an average of 25 units and a standard deviation of 2 units, using the algorithm ‘average + s × NORM.S.INV(RAND())’ [9]. The average and standard deviation of an increasing number of draws were calculated. The corresponding real case occurs if a series of the same sample were measured in duplicates and their differences normally distributed.

2. Random values between 23 and 27 units (25 ± 2 units) were chosen as original observations, i.e. a rectangular distribution with a measurement interval of 4 units and a random difference between −5 and +5 added, i.e. a standard deviation of \( 5/\sqrt{3} = 2.89 \) units. The corresponding real case would be a measurement procedure where the results differ randomly within a defined interval.

3. The original measurements were simulated as a normal distribution. The replicates were made dependent and the difference between the original and replicate was normally distributed with a constant standard deviation, i.e. ‘s × NORM.S.INV(RAND())’ and no bias assumed. This mimics a homoscedastic profile.

4. The original observations were randomly chosen within the interval 10 to 30 units. The differences were randomly chosen between –3 and +3, i.e. the standard deviation estimated to \( 3/\sqrt{3} = 1.73 \) units, assuming a rectangular distribution. The repeatability was calculated from the differences between the replicates and the originals, expressed in relation to the average of the original and replicate.

5. The original series was a set of randomly chosen numbers between 10 and 30 units. The replicates were created as a function of the original and a randomly chosen proportionality factor between 1.005 and 1.025 and therefore dependent on the original observation. The relative differences between the original and the replicate were used to calculate the standard deviation. This mimics a heteroscedastic profile.

The target value was defined as that obtained by simulation of a sufficiently large number of observations. In the present study, a target value of 1.12 was obtained as the average of 100 simulations of 40 observations.

Based on these assumed situations, five sets of simulations were made, each comprising 50 samples. The Dahlberg formula was applied to two, three, four and so of pairs of the dataset. Results of standard deviations based on 2, 5, 10, 15, 20, 25, 30, 35 and 40 pairs were then tabulated and
repeated 100 times. The average of the variances which were calculated this way and their confidence intervals are shown in Figure 3, left panel.

To investigate the influence of bias on the calculations and to illustrate the difference between the Dahlberg formula and the 'expanded Dahlberg' formula, appropriate series of \(0.5 \times 10^6\) observations were simulated. The average of the 'original' series was set to 10 with a standard deviation of 1, thus defining the measuring interval. The random error to create the 'repeat' value was based on a normal distribution with an average of zero and a coefficient of variation of 2.5% or 5.0%. As a third series, the 'repeat' series was modified by adding a constant error ranging from 1 to 20% of the average of the original series of measurements. This represents a bias or systematic variation between the original and repeat measurements.

Results

Analysis of variance components using reference materials

Results of the ANOC are exemplified in Figure 2 using the results from the two calcium reference materials measured 5 \(\times\) 5 times. There was one outlier in the first series of the low concentration. This observation was not included in the calculations but the results are included for the sake of information. The results of ALT, ALP and Ca are summarized in Table 1. In all experiments the within series (repeatability) was the dominating source of uncertainty. In the series of low Ca-concentration the mean square of the reproducibility (between series) was lower than that of the within group for the Ca-measurements of the low concentration material and therefore not presented.

Reproducibility by IQC

The average and standard deviation were calculated by standard methods, across all results, of the daily measurements of control materials and shown in Table 2. Since the results were collected during an extended period of time, they were regarded as obtained under reproducibility conditions and in this case also representing intra-laboratory uncertainty or total imprecision.
Repeatability, pure reproducibility and total uncertainty were also calculated from the IQC observations. For the high concentration materials of all analytes and the low concentration material for calcium the $M_{b}$ proved smaller than $M_{w}$ and the total variance therefore equal to the repeatability. The variance across all observations was equal to that calculated by analysis of variance components.

**Repeatability imprecision estimated by duplicate measurements**

The P-ALT, P-ALP and P-Ca concentrations were measured as duplicates in 100 consecutive routine samples. Only the first value was reported to the clinic to be used in patient care. The duplicate results were analyzed by the Dahlberg formula and the ExpD [4,11], (Table 3). Since the

<table>
<thead>
<tr>
<th>ALAT μkat/L</th>
<th>ALP μkat/L</th>
<th>Calcium mmol/L</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average</td>
<td>0.74</td>
<td>2.32</td>
</tr>
<tr>
<td>Count</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Across (s)</td>
<td>0.013</td>
<td>0.020</td>
</tr>
<tr>
<td>Maximum</td>
<td>10.3</td>
<td>16.6</td>
</tr>
<tr>
<td>Minimum</td>
<td>0.1</td>
<td>0.5</td>
</tr>
<tr>
<td>s(X) (Dahlb)</td>
<td>0.016</td>
<td>0.023</td>
</tr>
<tr>
<td>%CV</td>
<td>1.76</td>
<td>0.90</td>
</tr>
<tr>
<td>Pure Reproducibility (s)</td>
<td>0.009</td>
<td>0.022</td>
</tr>
<tr>
<td>Intra-laboratory (s)</td>
<td>0.013</td>
<td>0.022</td>
</tr>
</tbody>
</table>

Table 2. The imprecision (standard deviation) calculated across all observations and using the method of analysis of variance components from 2 to 6 daily IQC measurements of control material, comprising 83 observations.

The five simulated models differed with respect to the design of the measuring interval and the relation between the original and repeated results. The original results were either normally distributed with limited tails or taken from a rectangular distribution. The repeated results were either obtained by the addition of a random number for a normal distribution or obtained by multiplication with a random, reasonable factor. Therefore, the original and repeat results were related in a defined fashion. As expected, the estimated standard deviation approached the target value as the number of included pairs increased and its variance and confidence interval decreased. Beyond 25 pairs, the standard deviation stabilized and a further decrease of the confidence interval was small. Similar results were obtained for all the simulation models (Figure 3).

**Simulations with non-random variations**

The large number of simulated data underpinning the Table 4 provides robust numbers for illustration of the influence of a non-random difference between the first and repeat observations. The first measurements were a series of $0.5 \times 10^6$ values drawn from a normal distribution. The series of repeat results was obtained by adding a random number to the original. The averages of the first and repeat measurements were thus identical and the average difference was zero (Table 4). In a third series, a constant was added to the repeat series and therefore the difference between these series and the original is the bias. If the standard deviation was calculated using formula (Equation (3)) after addition of the bias, it amounted to 2.0304 whereas if formula (Equation (5)), ‘extended Dahlberg’ was used the same estimate was obtained as for the un-biased series. The simulated series were also subjected to evaluation by ANOC. The repeatability was the same as that obtained by the Dahlberg formula.

| s(X)(Dahlb) | 0.0016 |
| %CV(X)     | 0.326  |
| Maximum difference | 0.10 |
| Average difference | 0.0004 |
| Student tde | 0.17 |
| $p$ (2-tail) | 0.863 |
| s(X)(ExpD)  | 0.016  |

The repeatability (bold) was calculated by Dahlberg’s formula and the expanded formula (ExpD). s(X) and %CV(X) represent the absolute and relative Dahlberg uncertainty, respectively.

**Table 3. Properties of duplicate measurements of patient samples.**

The necessary sample size

The five simulated models differed with respect to the design of the measuring interval and the relation between the original and repeated results. The original results were either normally distributed with limited tails or taken from a rectangular distribution. The repeated results were either obtained by the addition of a random number for a normal distribution or obtained by multiplication with a random, reasonable factor. Therefore, the original and repeat results were related in a defined fashion. As expected, the estimated standard deviation approached the target value as the number of included pairs increased and its variance and confidence interval decreased. Beyond 25 pairs, the standard deviation stabilized and a further decrease of the confidence interval was small. Similar results were obtained for all the simulation models (Figure 3).

**Table 4. Summary of the simulations of a normally distributed set of $0.5 \times 10^6$ simulated results (Orig) from an average of 10 and a standard deviation of 1 and a simulated repeated series with an assumed average difference of 0.25 times a random normality probability function (Rep).**

<table>
<thead>
<tr>
<th>Set values</th>
<th>10</th>
<th>1</th>
<th>0.25</th>
</tr>
</thead>
<tbody>
<tr>
<td>x</td>
<td>9.9993</td>
<td>9.9992</td>
<td>-0.0001</td>
</tr>
<tr>
<td>Sumsq(x−x̄)</td>
<td>5491200</td>
<td>50521745</td>
<td>31249</td>
</tr>
<tr>
<td>Stand dev (s)</td>
<td>0.9993</td>
<td>1.0306</td>
<td>1.0306</td>
</tr>
<tr>
<td>Var Dahlberg</td>
<td>0.0313</td>
<td>0.0313</td>
<td>0.0313</td>
</tr>
<tr>
<td>Var Expanded</td>
<td>0.0313</td>
<td>0.0313</td>
<td>0.0313</td>
</tr>
</tbody>
</table>

A non-random difference, equal to zero and 2 (20%), right panel illustrates how a traditionally calculated SD$_{Dahlb}$ differs from that using the expanded Dahlberg formula, which gives an SD$_{ExpD}$ devoid of an influence of the average of the bias. Finally, calculated quantities are in bold. The negligible effect of a random error between the original and repeated measurement is clearly shown as the variance of the repeats as the sum of the variances of the difference and original series (see text).
The absolute difference between a standard deviation estimated by the Dahlberg formula and the extended version of the Dahlberg formula is linear in relation to the relative size of the bias and independent on the measuring interval. The relative difference between the standard deviation estimated by the two formulas is related to the size of the non-random variation (Figure 4, right).

Discussion

Different strategies to determine the measurement uncertainty in a medical laboratory were investigated. The uncertainty is linked to the measurement method as a method uncertainty and this concept is commonly recognized as a characteristic that should meet relevant quality goals. The standard uncertainty is by definition expressed as a standard deviation and therefore related to the dispersion of repeated measurements. However, in laboratory medicine samples may be from the same patient and the object of the measurement is to verify a difference between the samples. Since the samples may have been collected under different circumstances, the laboratory needs to consider establishing different measurement uncertainties for use in appropriate situations. Monitoring an acute clinical situation e.g. a potential myocardial infarction with Troponin is performed under repeatability conditions whereas for instance screening for a disease by comparing a result with a reference value or monitoring a chronic disease are considered as performed under reproducibility conditions. The analytical sensitivity i.e. ability to verify a difference between results, increases as the uncertainty decreases. Consequently, there is not a single analytical quality goal and no single optimal method for expressing and monitoring analytical performance appropriate in all clinical contexts. In most cases, the repeatability uncertainty is less than the reproducibility uncertainty which is influenced by numerous factors. That does not exclude the possibility that the 'pure' reproducibility may be smaller than the repeatability. It is important for the laboratory to identify all sources of uncertainty to control causes of variability, minimize uncertainty and initiate corrective actions and differentiation of the uncertainty, e.g. recognizing the repeatability and reproducibility are valuable pieces of information.

Repeatability conditions are unchanged conditions i.e. using the same procedure, reagents, calibration, temperature etc. This is also known as the within series conditions. The within series uncertainty or simply 'repeatability' can be determined by repeated measurements of a control material under suitable constant conditions. If these 'within series' measurements are repeated, e.g. in a common IQC procedure, the repeatability and the reproducibility can be calculated using an ANOVA, analysis of variance components. The relevance of this information applies to the concentration of the used material only and is not necessarily representative for the entire measuring interval. Furthermore, the variance of variance components estimates depend on the underlying distributions [12,13]. The Dahlberg formula is derived differently and allows calculating the repeatability from duplicates of many samples of different concentrations [14]. A key question is how many replicates would be necessary to achieve a robust estimate of the uncertainty. Through simulation, we estimated the minimal number to 25 repeated samples. This was verified in original measurements of patient samples.

The Dahlberg formula (Equation (4)) assumes a negligible variation of the uncertainty of the tested results or that the differences between the first and repeat observation are random, i.e. normally distributed. If this is not the case there is a systematic difference between the first and repeat measurement, i.e. a bias. This can be handled by the expanded Dahlberg formula which is the sum of squares of the differences between observed differences and the average of all differences. If there is no bias, the results of the formulas will coincide and it is therefore a good practice to apply both, which allows the bias to be quantified.

If repeatability conditions are not fulfilled, reproducibility conditions exist i.e. one or more measurement conditions have been changed. By necessity, the reproducibility...
‘includes’ repeatability and to calculate the correct intra-laboratory variance, the ‘pure’ reproducibility should be calculated. The total or intra-laboratory variance can then be calculated as the sum of the two variances. This value can be compared to the variance calculated across all values. Particularly if the pure reproducibility is of a noteworthy size, the variance calculated across the values underestimates the intra-laboratory variance.

The principle of ANOC is a powerful tool for investigating the performance of a measurement method. It requires a special experimental setup to ensure that both repeatability and reproducibility conditions are studies. The technique has been adopted by the CLSI EP15 to verify the quality claims of a manufacturer [5]. Their setup is limited to 5 × 5 observations with the warning that the power of the used degrees of freedom is not sufficient to establish, only verify a method variance. A similar approach is used in the CLSI EP5 [15]. The technique has also been used in real time monitoring the comparability and performance of several laboratories which allows identifying problems within the individual laboratories and between those participating [16].

The repeatability imprecision was estimated by the Dahlberg formula (Table 3). There are numerous situations where, and reasons why, non-random differences between repeated measurements occur, e.g. carry over, reagent decay i.e. the experimental conditions are not truly repeatable. It is essential to design an experimental model which minimizes any systematic difference due to laboratory factors, i.e. bias, between the first and repeat measurements. In the present method and modelling study we have not considered any influence of pre-analytical variation. In large enough series the averages of the original and repeat observations coincide. This indicates a technique to estimate any bias either by simply comparing the averages of the series or applying the Student’s paired t-test. It is possible to estimate the bias also by applying and comparing the extended Dahlberg formula to the data. In the present study the two variances were equal and thus no bias between the two results. This was verified by the Student’s test for paired observations which supports the null hypothesis that there was no difference between the first and second results (Table 3). An overestimation of the repeatability imprecision by the Dahlberg formula was reported by Springate [17] which could possibly have been caused by a non-random difference.

The expanded Dahlberg formula compensates for the non-random error and estimates the within series variation correctly (Table 4) in the presence of a bias between duplicates. It was shown in the simulation that the original Dahlberg formula will overestimate the repeatability in the presence of a bias, i.e. include the bias in the uncertainty estimate. The ANOC is liable to the same effects. The relative difference between the two estimates increases almost exponentially up to a relative bias of about 1.5% and subsequently increases almost linearly (Figure 4, right panel), whereas the square root of the difference between the calculated squared Dahlberg and Expanded formulas increases linearly. The magnitude of the ratio between the two estimates is also related to the measuring interval.

Conclusions

Analytical goals are conventionally defined according to a hierarchy agreed by the laboratory professions. The degree to which laboratories and accreditation bodies meet the requirements is often judged by evaluating measurements of control materials without considering their commutability with patient samples. Duplicate measurements of patient material for establishing and monitoring performance can be an option.

The analytical quality goals commonly do not differentiate between the repeatability and reproducibility and it is less obvious how the laboratories should measure the real uncertainty of measurements. As investigated in the present study, the uncertainty can be determined under repeatability and reproducibility conditions and it is argued that both are important for the laboratory and for the end-users of the results.

The experimental design to determine the quantities is critical and the clinical use of the information depends on the clinical situation. Evaluation of biochemical markers for a fast-changing disease should preferably rely on the repeatability whereas a slow process or trend should rather be based on the reproducibility or combined intra-laboratory variation. The laboratory can use the detailed information of variance components for quality control and rational corrective actions. In general, the repeatability uncertainty is smaller than the reproducibility or intra-laboratory uncertainty and therefore offers a more sensitive diagnostic tool.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This study was supported by the Karolinska university laboratory and the County council of Östergötland.

References


